Discovering new verb-preposition combinations in New Englishes

Schneider, Gerold ; Zipp, Lena

Abstract: The grammatical description of New Englishes is a relatively young field but at the same time one that benefitted much from recent developments in corpus linguistics. Standard reference corpora such as the International Corpus of English (ICE) have made it possible to research grammatical phenomena even in smaller outer circle varieties of English. In the field of grammar, innovations typically start out at the intersection of grammar and lexis. We investigate verb-preposition combinations in four corpora of first and second language varieties of English, among them the preliminary version of the written component of ICE Fiji. Our focus is on what has been termed 'new prepositional verbs' (cf. Mukherjee 2009, Nesselhauf 2009), i.e. novel combinations of verbs and prepositions. We compare a manual and a semi-automated approach to the study of new verb-preposition combinations. The manual approach consists of a surface search for prepositions followed by a careful manual filtering process. The semi-automated approach is a corpus-driven investigation using parsed corpora and detecting variation-specific prepositional collocations. Typically, the advantage of manual searches is that precision is very high; the disadvantage is that the investigation is time-consuming and recall can be incomplete, because the scope of investigations may have to be restricted. The advantage of automatic, parse-based methods is that they are fast and corpus-driven, which may increase recall; the disadvantage is that error-rates are high, which seriously affects precision. We discuss similarities and differences in the results of the two approaches and show examples of new verb-preposition combinations from ICE India and ICE Fiji that the two approaches deliver. We conclude that both methods validate, but also complement each other.

Discovering new verb-preposition combinations in New Englishes

Gerold Schneider and Lena Zipp
English Department, University of Zurich

Abstract

The grammatical description of New Englishes is a relatively young field but at the same time one that benefitted much from recent developments in corpus linguistics. Standard reference corpora such as the *International Corpus of English* (ICE) have made it possible to research grammatical phenomena even in smaller outer circle varieties of English. In the field of grammar, innovations typically start out at the intersection of grammar and lexis. We investigate verb-preposition combinations in four corpora of first and second language varieties of English, among them the preliminary version of the written component of ICE Fiji. Our focus is on what has been termed 'new prepositional verbs' (cf. Mukherjee 2009, Nesselhauf 2009), i.e. novel combinations of verbs and prepositions.

We compare a manual and a semi-automated approach to the study of new verb-preposition combinations. The manual approach consists of a surface search for prepositions followed by a careful manual filtering process. The semi-automated approach is a corpus-driven investigation using parsed corpora and detecting variation-specific prepositional collocations. Typically, the advantage of manual searches is that precision is very high; the disadvantage is that the investigation is time-consuming and recall can be incomplete, because the scope of investigations may have to be restricted. The advantage of automatic, parse-based methods is that they are fast and corpus-driven, which may increase recall; the disadvantage is that error-rates are high, which seriously affects precision. We discuss similarities and differences in the results of the two approaches and show examples of new verb-preposition combinations from ICE India and ICE Fiji that the two approaches deliver. We conclude that both methods validate, but also complement each other.

1. Introduction

1.1 Corpora for New Englishes

The detailed grammatical description of New Englishes is a comparatively recent trend. Previous descriptive approaches of grammatical phenomena in second language varieties of English (ESL) relied largely on anecdotal evidence (see e.g. Foley 1988, Bautista and Gonzales 2006), or at best on the (mainly manual) analysis of sociolinguistic data (see Schreier 2003) rather than on representative collections of text. The *International Corpus of English* (ICE) project began to compile standard reference corpora of first and second language varieties of English in the early 1990s, and regional components are now becoming available for many ESL varieties: ICE Philippines was released in 2005, ICE Jamaica came out in 2009, and ICE Fiji (amongst others) is currently being compiled (Biewer et al. 2010). This set of comparable corpora provides the basis for corpus-linguistic studies that complement sociolinguistic investigations; it makes representative data available that ranges across various text-types, including the upper end of the stylistic spectrum. Furthermore, its matching design allows for comparative studies of a quantitative nature.

1.2 Corpus-based and corpus-driven approaches

Overall, recent corpus-based descriptions of ESL varieties (see Sand 2004, Schneider 2004 or Sedlatschek 2009) have been conducted on orthographic, i.e. not annotated, corpora, as work on tagging and parsing the ICE components is still under way. As a result, these descriptions have to rely on more or less sophisticated searches based on lexical items. It may be possible, however, to arrive at a partly corpus-driven description of grammatical phenomena in New Englishes: Mukherjee and Hoffmann (2006), for example, make use of tagged web-derived collections of text, and Xiao (2009) employs a tagger on five ICE corpora to conduct a multidimensional register analysis. In our approach, we use more richly annotated corpus material: The required corpora are annotated syntactically and our aim is to explore in how far this annotation may yield useful information for the description of New Englishes. In turn, the New Englishes databases might be exploited in fine-tuning the annotation tools to the structural challenges that second language varieties of English present; this very approach is described and tested on selected phenomena in Schneider and Hundt (2009). In the present study, we compare the traditional method of carefully analysing orthographic corpora manually with a corpus-driven approach based on automatically parsed corpus data; we focus on the case of lexico-grammatical phenomena in the verb phrase.

1.3 Lexico-grammar

Studies investigating patterns located at the lexis-grammar interface can in principle draw from the two main characteristics of the field in question. Starting a search query from the purely lexical end soon reveals that all content words express semantic differences and are very specific to the topics that happen to be discussed in the selected texts. As described by Zipf's law, most content words are rare, which leads to a sparse-data-problem even when using large corpora. If corpora of appropriate size are used, the results are dominated by regional differences such as place names and semantic differences. Function words are a more promising starting place for lexical search queries: Because they do not express semantic concepts directly, they are less affected by regional or semantic differences. Furthermore, while being comparatively frequent, they also form closed lists, which facilitates manual search, especially in languages with as little morphology as English. We will explore this point of departure in section 2.

When starting investigations at the grammatical end, for example by comparing frequencies of part-of-speech tags, phrase types, or grammatical relations, it can be noticed that regional differences are relatively small while genre differences are often bigger (Biber, Conrad and Reppen 1998). While not impossible (and a crucial task for future research), it will be very difficult to disentangle genre differences from regional variation. Variational differences are often too subtle to leave a visible impact in frequency counts. In fact, the vast majority of sentences in e.g. ICE India or ICE Fiji could just as well have been produced by a British or American speaker, there is nothing 'unusual' in them.

While one-dimensional investigations of only the lexicon or of only the grammar may lead to limited success, it has been observed that the crucial variationist differences happen in the interaction of lexis and grammar. Schneider (2004: 229) for example states that in World Englishes,

> distinctive phenomena tend to concentrate at the interface between grammar and lexicon, concerning structural preferences of certain words (like the complementation patterns that verbs allow), co-occurrence and collocational tendencies of words in phrases, and also patterns of word formation.

It will thus be revealing to investigate the lexical material that is used in syntactic relations. As a first approximation to words in syntactic relations, one can investigate surface word or word-tag sequences. For example, investigating trigrams that are frequent in ICE India but absent in the one-hundred-times larger *British National Corpus* (BNC) leads to the list given in figure 1, after filtering trigrams containing proper names and punctuation. Besides text selection, Indian features like archaic spellings (*now a days*), formal language (*the honourable minister*), unusual verb complementation with prepositional phrases (*is called as*), and written numbers (*sixty-six and half*) appear in this list. Examples that show the unusual verb complementation trigram *is called as* are:

(1) A substance which is helping in chemical reaction is called as a reagent. (ICE-IND:S1B-004)

(2) Thus the intermediate state between crystaline and isotopic state is called as the mesophase or liquid crystals. (ICE-IND:W1A-020)

Trigram	f(ICE-India)
now_RB a_DT days_NN	42
special_JJ P_NN P_NN	35
canvassed_VBN before_IN this_DT	32
statement_NN was_VBD recorded_VBN	28
learned_JJ special_JJ P_NN	28
is_VBZ called_VBN as_IN	27
scene_NN of_IN offence_NN	26
the_DT honourable_JJ minister_NN	23
for_IN grain_NN yield_NN	22
the_DT learned_JJ special_JJ	21
in_IN the_DT cyclone_NN	19
delay_NN in_IN reply_NN	18
best_JJS feature_NN film_NN	18
avoid_VB delay_NN in_IN	18
small_JJ circle_NN to_TO	17
of_IN solid_JJ wastes_NNS	17
general_JJ body_NN meeting_NN	17
evidence_NN of_IN P_NN	17
feature_NN film_NN in_NN	16
crores_NNS of_IN rupees_NNS	16
in_IN the_DT nodules_NNS	15
has_VBZ also_RB canvassed_VBN	15
sixt-six_NN and_CC half_NN	14

Figure 1. Unusual trigrams in ICE India.

Is called as in the examples above belongs to a field of lexico-grammar that has been noted to hold great potential for studies of variety-specific usage patterns: verb complementation. For example, Olavarría de Ersson and Shaw (2003: 138) state that “Verb complementation is an all-pervading structural feature of language and thus likely to be more significant in giving a variety its character than, for example, lexis.” A number of previous studies have focussed on the variability of verb-particle combinations (with both prepositions or adverbial particles) in particular (see Mukherjee and Hoffmann 2006, Mukherjee 2009, Nesselhauf 2009, Zipp forthcoming). All of these studies investigate the occurrence of ‘new prepositional verbs’, i.e. novel combinations of verbs and prepositions that are triggered by a process of ‘semantico-structural analogy’, “a process by means of which non-native speakers of English as a second language are licensed to introduce new forms and structures into the English language because corresponding semantic and formal templates already exist in the English language system” (Mukherjee and Hoffmann 2006: 166-167). In most cases, the result of this process is a verb-particle combination with a redundant, i.e. additional preposition attached to a verb that does not usually combine with a particle (this is the classic case investigated in most of the previous studies). However, other divergent types could be the following: cases of different preposition, missing preposition, or un-idiomatic usage of existing verb-particle combinations. The first type is also included in the analyses here: verb-particle combinations in which we find a different preposition than the ones that are codified. The second type, missing preposition, could not be

investigated by our research method. Schneider (2004) searches for a small set of specific verbs. While this allows the retrieval of missing particles, the complete set of verbs would be forbiddingly cumbersome. To complement Schneider (2004) we use a more corpus-driven all-inclusive method here. The third type however, un-idiomatic usage of existing verb-particle combinations, is addressed in the manual analysis in section 2. Regarding combinations of verbs and prepositional particles, traditional grammars distinguish between phrasal, prepositional and phrasal-prepositional verbs. For the investigation of verb complementation in this study, however, we purposely leave the distinction between preposition and verbal particle underspecified. All verb-preposition constructions are included, irrespective of whether they are specified or unspecified, continous or discontinuous. The manual approach only includes complements, whereas the automatic approach also delivers adjuncts.

1.4 Data

The data used for both methodological parts of this study comes from the same set of varieties of English: Fiji English, Indian English, New Zealand English and British English. This selection is justified by the original research object of previous studies, Fiji English, and its geographical, historical and cultural relations to India, New Zealand and Great Britain. The Fiji, Indian and New Zealand data are all part of the *International Corpus of English* (ICE) project; for the purposes of the present study, however, it was necessary to work with different datasets due to the methods we report on: The manual method is only concerned with the respective written parts of the corpora; ICE Fiji is still in the process of compilation at the moment of writing, and therefore represented by a part of the written subcorpus only. For the corpus-driven parsing method, all corpora were automatically parsed; this includes the Fiji corpus and the complete (i.e. spoken and written) regional components for Indian and New Zealand English. For reasons of corpus size, the basis of comparison for the parsing method is the written BNC corpus with approximately 90m running words (see Table 1).

Manual method				
ICE written components	Fiji	India	NZ	GB
number of 2,000 word files	140	200	200	200
Parsing method				
ICE	Fiji	India	NZ	BNC written
number of 2,000 word files	140	500	500	90m words

Table 1: Corpora used

2. Manual method

This section reports on the processes that we followed and the results that we achieved on the task of investigating new verb-particle combinations in untagged corpora of New Englishes. All technical levels of this traditional corpus-linguistic study, which is based on lexical search queries, serve as the matrix against which the corpus-driven, parser-based method described below (section 3) is evaluated. As mentioned above, lexico-grammatical phenomena are claimed to be very good indicators of variety-specific structural nativisation. From a formal point of view, the grammatical elements of lexico-grammatical phenomena are represented by function words, including e.g. determiners (see Schneider and Hundt 2009), auxiliaries, modal verbs (see Biewer 2009), complementizers, and prepositions, which we will investigate here. In the specific case of verb complementation by particles, verb lexemes in all their inflectional forms combine with prepositions, i.e. function words. Prepositions are members of a closed class and not subjected to morphological operations; they can thus easily and exhaustively be found with simple lexical searches.

The first step of the manual investigation was therefore a lexical surface search for prepositions. For reasons of efficiency, the search queries were limited to the five most productive particles in verb-particle combinations (see Villavicencio 2006), *up, out, down, off* and *away,* and two prepositions that have repeatedly been claimed to be productive in the formation of new particle verbs in New Englishes, *into* and *about*. This was followed by a careful manual filtering process, as a result of which all instances were eliminated in which the prepositions did not occur within the verb phrase, or constituted false positives (e.g. due to linebreaks, within editorial comments, or added with the help of corrective mark-up). Whenever a verb-preposition combination was not recorded in the following sources, it was considered as 'unrecorded': The *Collins COBUILD Phrasal Verbs Dictionary* (2002), the *Collins COBUILD Advanced Learner's Dictionary (Resource Pack CD)* (2003), the *Oxford English Dictionary Online* (2009) and, in selected cases, the Internet by means of Google search in selected cases.

	ICE Fiji	**ICE IND**	**ICE GB**	**ICE NZ**
V + up	38 (11)	31 (13)	5 (2)	5 (2)
V + out	14 (4)	31 (13)	-	9 (4)
V + down	10 (3)	2 (1)	-	5 (2)
V + off	7 (2)	24 (10)	-	2 (1)
V + away	3 (1)	24 (10)	11 (5)	16 (7)
V + into	52 (15)	24 (10)	9 (4)	9 (4)
V + about	28 (8)	17 (7)	2 (1)	7 (3)
total pmw (raw)	**152 (44)**	**153 (64)**	**27 (14)**	**53 (23)**

Table 2: Distribution of unrecorded verb-preposition combinations across varieties per million words and (raw)

Table 2 shows the total number and normalized distribution of unrecorded verb-preposition combinations in each of the written subcorpora used for the manual analysis. The difference between the total number of unrecorded combinations is statistically significant at the $p<0.001$ level (chi-square contingency test, d.f.=3). Indian and Fiji English exhibit the highest number of innovations, followed by New Zealand English and British English. Across varieties, the prepositions *up, away, into* and *about* are consistently used in new verb-preposition combinations. From a variety-specific perspective, the first language varieties (GB and NZ) exhibit the greatest productivity in combination with the preposition *away*, which in most cases is used to add an aspectual dimension of continuity to a procedural verb (with the particle establishing the notion of 'persistent action', see Quirk et al. 1985: 1162). On the other hand, the prepositions *into, up, about* and *out* are used most often in unrecorded verb-preposition combinations in the second language varieties under observation. Note that unrecorded verb-preposition combinations were found in all corpora, despite the limited data size and the partly lexical nature of the phenomena investigated.

In order to shed light on the types of verb-preposition combinations that were detected on the basis of the manual method, we will now present a selection of examples. As mentioned in section 1.3, there are four types of possible divergence in verb-particle combinations, of which three can be investigated with the help of the manual method: redundant particle, different particle, and un-idiomatic usage.

2.1 Redundant particle

Examples 3 to 5 show four instances in which a redundant preposition is added to a simple verb. This process of creating 'new prepositional verbs' by analogy to existing, semantically related particle verbs is described in detail by Mukherjee (2009) and Nesselhauf (2009), and discussed and applied by Zipp (forthcoming). Below, we present a range of typical examples: The combination *cope up with* has been noted before in the context of many New Englishes, *explaining about* and *discussing about* belong to a relatively homogenous group of disquisition verbs combining with this preposition (such as *talk about* or *speak about*), and *listed down* might be triggered by analogy to the verb *put down*.

(3) As a result some or nearly most women in the world have now turned to becoming prostitutes in order to cope up with poor living standard, they may be experiencing. (ICE-FJ:W1A-016)

(4) First, I would be explaining about the gender inequality, which often leads to the high incidence of poverty amongst women, which is what I would be discussing about in the second part of this essay. (ICE-FJ:W1A-016)

(5) Adi Asenaca said an Asian Development Bank poverty participation survey listed down forms of poverty in the country and her ministry was following up on the recommendations. (ICE-FJ:W2C-013)

2.2 Different particle

Sentences 6 to 11 are examples from the Fiji and Indian data in which unusual prepositions are used in combination with various verbs; a common phenomenon is the use of *off* for *of,* and *into* instead of *in*. The former may also arise from a typing mistake, but the frequency of its occurrence renders such an interpretation unlikely. It might be argued that the distinction between these prepositions is comparatively fine-grained and thus a pre-determined point of confusion in English. However, we do not aim at explaining the motivation of the phenomena we describe here; further research on the cognitive processes linked to the semantic perception of these two sets of prepositions will have to be undertaken.

(6) One of the side effects of alcohol is that it rids our body off nutrients, and the reason we feel like a truck has rolled over our heads is because we need vitamins to function. (ICE-FJ:W2D-012)

(7) We have allowed racism to manifest itself into the education system directly or indirectly through our actions or through the examples we have set to our students as role models. (ICE-FJ:W2B-007)

(8) Some of these are; women involving themselves into prostitution, selling their infants, migration across world, low standard in society. (ICE-FJ:W1A-018)

(9) In some situations, however, waste can be a big health hazard and must be disposed off properly, for example by sanitary land fill. (ICE-IND:W2A-031)

(10) Raju's work has eased out a bit. (ICE-IND:W1B-014)

(11) This resulted into a deep sense of growing loneliness which affected the individual life. (ICE-IND:W2A-005)

2.3 Un-idiomatic usage

The last type of new verb-preposition combinations occurs in two sub-types: combinations that were used in contexts that do not match the interpretations given by dictionaries (examples 12 to 14), and combinations in which an existing particle verb is used where the simple verb would be the more appropriate choice (examples 15 to 18). These instances of un-idiomatic usage benefit from semantic evaluation of the context in which the verb occurs; the combination itself is recorded and thus difficult to detect by automatic retrieval methods. At best, they could be automatically detected based on the increased frequency of the particular verb-preposition combination.

(12) By the by, we're looking for a media person, someone who can front up to the hacks without crumbling. (ICE-FJ:W2F-016)

(13) Tukania picked up competitive football in 1980 and a year later forced his way into national coach late Billy Singh's South Pacific Games squad. (ICE-FJ:W2C-019)

(14) When the migrant races want to dominate us economically and now politically, through the 1997 Constitution, even though we have a higher population distribution of 51 per cent, the so-called democratic system does not stack up for our rights and differences. (ICE-FJ:W2B-012#34:1)

(15) Coming over to play Fiji is an experience no one can rob them of, it's not about the win but the exposure and the pride to play up against one of the world's best is all that counts," she said. (ICE-FJ:W2C-007)

(16) Indian women were mostly dressed up in sarees. Even the woman indentured labourers came to work on plantations in sarees. (ICE-FJ:W2B-008)

(17) Exceed that and it will be hello hangover the following day. Drink up water in betweens, it will fill you up very quickly, and the many trips to the bathroom will flush out the alcohol. (ICE-FJ:W2D-012)

(18) The women participation in labour-force, more than doubled up between 1960-90. (ICE-IND:W2A-005)

2.4 Results

The manual method described above consists of combining a lexical surface search for the function words in multi-word verbs with a manual filtering process and analysis of the hits. It produced a significant number and range of results, i.e., a variety of new particle-verb combinations from all national varieties of English under observation. It has to be stressed, however, that this method is only concerned with detecting possible new combinations, not with assessing their status. Whether a new combination finally enters the lexicon of a particular variety of English and becomes standardised will have to be investigated on the basis of larger amounts of data, or by follow-up investigations of a diachronic nature. For the time being, it cannot be ruled out that a considerable number of new combinations are potential nonce formations (see example 19 and 20).

(19) Manju was in Goyle, a nearby village, and Manju's parents were always coconut wirelessed about her health and happenings. (ICE-FJ:W2F-013)

(20) One of the creepers had tubers the size of large turnips that we had to tomahawk out. (ICE-NZ:W1B-008)

However, the value of this analysis for determining potential starting points for further investigation of structural nativisation cannot be denied. Along the same line of argumentation, we refrain from judging whether the phenomena we report are indeed instances of variety-specific usage or performance errors. We believe that this distinction is above all a question of ideology; phenomena that are interpreted as instances of structural nativisation by variationist linguists are often seen as learner errors or substratum interference within the paradigm of second language acquisition, or slips of the tongue in the field of psycholinguistics. In the future, studies based on larger amounts of text will hopefully give a clearer picture of the respective frequencies; slips of the tongue will remain singular or at least rare occurrences, while structural nativisation phenomena will report more hits.

3. Parsing method

3.1 Using parsers for descriptive linguistics

Parsing technology has made considerable advances recently, opening new perspectives for descriptive linguistics. Van Noord and Bouma (2009: 37) state that "[k]nowledge-based parsers are now accurate, fast and robust enough to be used to obtain syntactic annotations for very large corpora fully automatically." We apply parsed corpora as a new resource for linguists. Automatically parsed treebanks, also called tree jungles, have been used for e.g. Danish (Bick 2003) and French (Bick 2010). No treebanks for English regional varieties or World Englishes exist yet. In this situation, automatically parsed corpora can be used as a stopgap to Treebanks. We have parsed the available ICE corpora and many other large corpora like the BNC using a dependency parser (Schneider 2008).

The semi-automated corpus-driven approach using parsed corpora is described in detail in Schneider and Hundt (2009). Here we apply it to the detection of variety-specific prepositional collocations. Advantages of (semi-)automatic, parse-based methods are that they are fast and corpus-driven, which may increase recall. A disadvantage is that error-rates are still relatively high in automatic parsing, which seriously affects precision. The small size of the ICE corpora poses an additional challenge: The detection of rare collocations is particularly difficult due to the low counts.

We have used a probabilistic dependency parser, Pro3Gres (Schneider 2008), which is quite fast (the BNC parses in 24 hours) and which has been evaluated on several genres and varieties (Haverinen et al. 2008, Lehmann and Schneider 2009). The grammar can be adapted manually to genres and varieties. We have used the same grammar on all ICE corpora, in order not to risk adding skews. The parser is suitable for parsing different varieties of English, because it is robust and because its output has been evaluated on a number of English varieties (Schneider and Hundt 2009). For example, it does not enforce subject-verb agreement, it allows zero-determiners everywhere, it uses statistical preferences instead of strict subcategorisation frames. This entails for example that non-ditransitive verbs can act as ditransitive, and that prepositional phrases with divergent prepositions get attached, a feature that we need for our investigation here. The parser outputs intuitive dependency relations. A subset of them is given in table 3. Verb-PP (prepositional phrase) attachment is expressed by the dependency relation *pobj*.

RELATION	LABEL	EXAMPLE
verb–subject	*subj*	he sleeps
verb–direct object	*obj*	sees it
verb–second object	*obj2*	gave (her) kisses
verb–adjunct	*adj*	ate yesterday
verb–subord. clause	*sentobj*	saw (they) came
verb–pred. adjective	*predadj*	is ready
verb–prep. phrase	*pobj*	slept in bed
noun–prep. phrase	*modpp*	draft of paper

noun–participle	*modpart*	report written
verb–complementizer	*compl*	to eat apples
noun–preposition	*prep*	to the house

Table 3. Important dependency relations that are output by Pro3Gres

3.2 Parser evaluation

An evaluation of the performance on subject, object PP-attachment and subordinate clause relations, using the 500 sentence GREVAL gold standard (Carroll et al. 2003), is given in table 4. Compared to others parsers, these rates are competitive (Schneider 2008). While some of the performance values may appear low at the first sight, the following facts alleviate the impact of errors: first, only precision errors indicate a wrong assertion, while recall errors entail that an instance has been missed, the signal remains unaffected. Second, errors are largely unsystematic, which means that the signal is weakened but skewed much less than by the error rate. Third, PP-attachment performance on complements (which is what we mainly need for this application) is better than on adjuncts.

	Subject	**Object**	**PP-attachment**	**clausal**
Precision	92.3% (865/937)	85.3% (353/414)	76.9% (702/913)	74.3% (451/607)
Recall	78.0% (865/1095)	82.5% (353/428)	68.6% (702/1023)	61.7% (451/731)

Table 4. Performance on the GREVAL gold standard corpus

In order to assess if performance is affected by variational differences, we have manually evaluated 100 random sentences from ICE GB and ICE Fiji and found similar performance (Lehmann and Schneider 2009, Schneider and Hundt 2009).

3.3 Detecting rare PP-collocations semi-automatically

As our method for detecting PP-collocations, we use the “surprise about” finding specific verb-preposition or verb-particle combinations. We use O / E (Observed / Expected) as measurement. We decided to use O / E instead of the t-test or log-likelihood, which are frequently used for the detection of collocations (see e.g. Evert 2009) for the following reasons: First, O / E is a measure of surprise, not of statistical significance. Collocation significance does not directly correspond to collocation strength, a measure of surprise may serve as a better proxy to measuring collocation strength. Second, O / E has the characteristic that it tends to give particularly high scores to rare events, which is beneficial for our purpose, as many of the new verb-PP combinations which we are investigating are very rare. In fact, they are often too rare to reach statistical significance. As considerable manual interaction is needed in our approach, manual validation of the suggestions made by the computer replaces the need for statistical significance. Third, O / E has been shown to work well for rare collocations, particularly if relatively clean data is used. Lehmann and Schneider (2009) use parsed data from the BNC and other large corpora to detect PP-collocations with O / E. While windows-based methods using O / E typically report a large amount of garbage in the top-ranked positions, O / E on parsed data delivers considerably better results (Lehmann and Schneider 2009). Windows-based methods (e.g. Stubbs 1995) are still commonly used for collocation detection. They use an observation window from N words before to N words after a key word (e.g. a verb) and count all words inside the window as co-occurrence. N is typically about 3. The distinction between different types of collocations (e.g. subject-verb, verb-object and verb-PP) is often left underspecified.

Windows-based methods typically lead to relatively many errors, both precision errors (false positives) and recall errors (false negatives). They suffer from precision errors due to the lack of implicit head extraction and due to the fact that words appearing close together are often not syntactically related. In the example sentence *We report on the Epstein Barr virus will spread* windows-based methods typically also report *report on Epstein* and *report on Barr* as verb-PP collocation counts due to the lack of head extraction. In the example sentence *The virus we reported on last week has dangerous consequences* windows-based methods typically report *week* and possibly, depending on N, *consequences*, as verb-PP collocation counts.

Recall is intrinsically low with windows-based methods because many of the dependencies appear further then N words away. Recall can be increased by increasing N, but at a forbidding cost of decreasing precision. We do not use the O / E measure directly, but we compare O / E obtained from an individual ICE corpus to O / E measures obtained from the BNC, in order to express how much more surprising the frequency of a verb-PP combination is in the ICE corpus under investigation, i.e. how much stronger a collocation is in an ICE corpus in comparison to the BNC. We compare to the BNC instead of ICE GB, because with ICE GB we experienced a serious sparse data problem. Very many verb-PP combinations, also many that are perfectly acceptable in British English, do not occur in ICE GB, while most of them appear in the BNC. As the ICE corpora are relatively small for our investigation, using a sufficiently large base of comparison can partly alleviate the sparse data problem.

We compare O / E measures by calculating a ratio. For the example of ICE Fiji, the formula is:

$$O/E\ ratio = \frac{O/E(Fiji)}{O/E(BNC)} = \frac{\frac{O(Fiji)}{E(Fiji)}}{\frac{O(BNC)}{E(BNC)}} = \frac{\frac{O_{Fiji}(R,w_1,w_2)\cdot N_{Fiji}}{O_{Fiji}(R,w_1)\cdot O_{Fiji}(R,w_2)}}{\frac{O_{BNC}(R,w_1,w_2)\cdot N_{BNC}}{O_{BNC}(R,w_1)\cdot O_{BNC}(R,w_2)}}$$

where N is corpus size, R is the verb-PP attachment relation (*pobj,* see table 3), w_1 the head verb, w_2 the preposition or verbal particle.

This formula assigns a value to the hundreds of verb-PP combinations that are seen in both corpora. The O / E ratio is above 1 if the collocation is more frequent in ICE Fiji (or whichever ICE corpus we apply), and below 1 if it is more frequent in the BNC. We are particularly interested in very high ratios, so we filter the list of verb-PP combinations, for example only to O / E ratio > 10 (i.e. at least ten times more surprising in ICE Fiji). The list thus obtained contains some surprising collocations and some collocations that are also acceptable and frequent in British English. The latter usually have high O / E values in the BNC, and due to coincidence, small corpus size, text selection, semantic content, etc. end up being more frequent in ICE Fiji. In order to filter them out, we also set a threshold on O / E values from the BNC: only O / E values below a certain threshold (we have used 3 in table 5), i.e. combinations that are not strong collocations also in the BNC are allowed.

We have also looked at verb-PP combinations that are present in an ICE corpus but absent in the BNC.

3.4 New verb-PP combinations in ICE Fiji

If we set the filter to O / E ratio > 10 and O / E in the BNC < 3 we get the list shown in table 5.

O / E ratio	Head	Prep	f (Fiji)	O / E (Fiji)	O / E (BNC)	manual inspection comment
14.4021	regard	to	7	41.9521	2.91292	serendipitous: he or she will be reading in regards to a bigger picture
14.616	cause	on	3	34.3407	2.34952	yes: The thought of how much anxiety he had caused on his parents ...
19.7136	stick	as	2	42.1458	2.1379	no
10.9451	pick	to	2	11.5253	1.05301	yes: allow me to pick my team to the world cup
33.9525	join	into	2	52.5526	1.54783	yes: Women by joining into these organisation benefit a lot
11.1615	involve	into	2	24.255	2.17311	yes: women involving themselves into prostitution
33.3689	include	into	2	65.2377	1.95505	yes: they have included rare ... species ... into the displays
22.3632	implicate	for	2	46.4807	2.07845	no
472.801	gather	upon	2	895.141	1.89327	yes, adjunct: upon evaluating the ... Education Act, it was gathered that ....
15.2663	explain	from	2	40.2206	2.6346	no, consistent parsing error
81.3601	engage	through	2	167.625	2.06028	no
31.246	concentrate	from	2	54.5852	1.74695	no
48.866	capable	in	2	14.2045	0.290684	yes, adjective: are capable in committing themselves to work
61.3927	arrive	into	2	43.9975	0.716656	yes: Megan Simpson is expected to arrive into the country

Table 5. Results from ICE Fiji. For O / E ratio > 10 verbs and O / E (BNC) < 3

The first column displays the O / E ratio as given in the formula. The second column contains the verb, and the third column the preposition in the PP-attachment relation. The fourth column, f (Fiji) reports how often the verb-PP combination is seen in ICE Fiji. Note that most of these values are very low, too low to reach statistical significance, which is one of the reasons why we have chosen O / E. In fact, we have also tested log-likelihood measures and obtained slightly worse results. Columns 5 and 6 show O / E from the two corpora. The last column contains our manual assessment ('yes' meaning this is a new Fijian verb-PP combination, 'no' meaning probably not) and an example for the cases where we have a typically Fijian verb-PP collocation. False positives are due to many different reasons; we have observed two as particularly frequent: first, consistent parsing errors. Second, the parser as we have used it here underspecifies the distinction between PP-argument and PP-adjunct, in order to increase recall: Unusual verb-PP argument combinations would hardly ever be recognised by the parser otherwise. This entails that frequent adjuncts, for example "concentrated ... from", which occurs repeatedly in scientific texts, appears in the list, or "upon ... it was gathered" which appears in judicial texts. We have decided to report the latter as it may be a candidate for very formal, seemingly slightly archaic expressions, which are generally more frequent in Asian English than in today's British English.

From a semantic perspective, many of these examples have been noted to display the "tendency to make the direction expressed in verbs of movement more explicit, even if this is already present in the meaning of the verb" (Nesselhauf 2009: 20, also see Zipp forthcoming). Some of these combinations of directional nature have been described before in selected New Englishes (e.g. Mukherjee 2009: 123, Nesselhauf 2009: 18, Sedlatschek 2009); the following are examples found in our data: *arrive into, include into, join into.* They can be seen as supporting the image of entering into a framed container or clearly framed status. Others may be seen as further specifying the verb meaning, for example *pick to* (restricting the meaning to *select,* which partly overlaps with *pick*), or as supporting the verb meaning, for example *cause on* (the preposition *to* is fairly neutral as in *give to, offer to*, the preposition *on* is negative, conjuring up *exert on, put on, looming on, impending on*).

As we can see in table 5, about half of the reported verb-PP collocations are false positives, so-called "garbage". Our approach does not intend to be fully automatic, and we are not aware of a fully automatic approach. Since the counts are too low to reach

statistical significance, and since the corpus linguist is interested in assessing and interpreting the results anyway, the manual filtering involved is usually acceptable and less work-intense than reading the whole corpus. In applications where the focus is on recall, less strict filters are used and a linguistic annotator has to filter more false positives. For example, with the very high O / E ratio > 40, but no O / E (BNC) threshold we get the list in table 6 from ICE-Fiji. We have selected thresholds that deliver interesting and particularly different results.

O / E ratio	Head	Prep	f (Fiji)	O / E (Fiji)	O / E (BNC)	manual inspection comment
381.198	reduce	amongst	7	1412.92	3.70652	no, consistent parsing error
89.4924	educate	than	3	763.081	8.52677	no
132.128	wrap	over	2	916.23	6.93443	no
169.581	tread	because	2	2287.58	13.4896	no
86.2943	renew	through	2	670.498	7.7699	no
121.186	renew	because	2	1715.69	14.1574	no
91.7332	poll	as	2	358.24	3.90523	no
51.3641	miss	without	2	474.255	9.2332	no
52.5289	know	behind	2	176.812	3.366	no, parsing error
50.3294	influence	towards	2	511.696	10.1669	yes: Leadership is defined as the ability to influence people towards the attainment of goals
472.801	gather	upon	2	895.141	1.89327	yes, see table 5
81.3601	engage	through	2	167.625	2.06028	no
130.402	enable	despite	2	5555.56	42.6032	no
124.367	award	over	2	732.984	5.89372	no
61.3927	arrive	into	2	43.9975	0.716656	yes, see table 5
429.654	anticipate	within	2	3381.64	7.87063	

Table 6. Results from ICE Fiji with O / E ratio > 40, but no O / E (BNC) threshold

While returning more false positives this list also contains a new finding, in which the preposition also seems to re-iterate the verb meaning : *influence someone towards something*. Repetitions of similar constructions affect the results. As several student essays in the ICE Fiji corpus are on the same topic, some combinations appear often: "*reduce poverty amongst women", "... are more educated than ... ",* and *"Through interpretation, tourist begins to engage"* appear in more than one student essay. In essence, these are sparse data problems.

3.5 New verb-PP combinations in ICE India

We have applied the same formula on other ICE corpora, particularly on other L2 corpora, where exonormative standardisation can be expected

For ICE India, using O / E ratio > 35 and O / E (BNC) < 3 as thresholds we obtain the results listed in table 7.

O / E ratio	Head	Prep	f (India)	O / E (India)	O / E (BNC)	manual inspection comment
80.6962	discuss	about	10	148.012	1.83419	yes: You come we will discuss about it.
51.3664	study	about	7	67.7127	1.31823	yes: Today we are studying about rotation and revolution of the earth.
705.33	advise	into	7	279.731	0.396597	no, consistent parsing error
39.8306	result	into	5	55.3685	1.3901	yes: This resulted into a deep sense of growing loneliness
78.7867	burst	of	5	234.214	2.97276	no
53.0517	arrest	from	5	59.374	1.11917	yes: five more terrorists were arrested from his home
93.5978	etch	at	3	147.232	1.57303	no
67.2343	withstand	to	2	139.353	2.07265	no
46.6381	significant	on	2	33.1642	0.711096	no
45.8399	nice	on	2	70.0133	1.52734	no
84.4974	line	of	2	120.453	1.42552	no
47.4123	land	into	2	102.124	2.15396	yes: Atul's tendency of worrying too much ... landed him into trouble
107.968	exciting	on	2	315.06	2.9181	no
214.685	benefit	out	2	128.156	0.596949	yes: So they'll benefit out of the faculty teaching

Table 7. Examples from ICE-India. For O / E ratio > 35 verbs and O / E (BNC) < 3

Again, there is a considerable amount of false positives that need to be filtered. The verb-PP combination *discuss about* is frequent in L2 and learner English as a whole (see above).

The parser output for one of the example sentences is given in figure 2. *Result into* is another example where the prepositional semantics is used to support the verb meaning. What is special in this case is that the existing phrasal verb *result in* which contains an opaque, semantically non-compositional particle *in* is rendered more transparent by the use of the preposition or particle *into*. This leads to a very similar construction which is probably ungrammatical in Standard English but has transparent semantics.

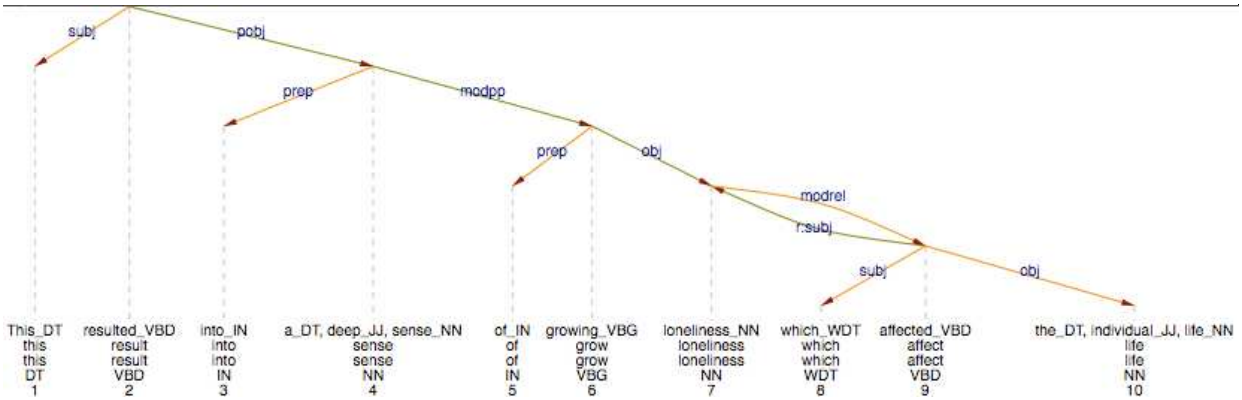

Figure 2. Automatic parse of *This resulted into a deep sense of growing loneliness, which affected the individual life*

For the experiments of section 3.4, we have hitherto used the entire ICE India, including the spoken part. This leads to a biased comparison, both compared to ICE Fiji in section 3.3 and to the manual method in section 2, where only the written sub-corpora of ICE India are used.

For O / E ratio > 8 and O / E (BNC) < 3 we get the short list given in table 8.

O / E ratio	Head	Prep	f (India)	O / E (India)	O / E (BNC)	manual inspection comment
17.7654	value	to	3	49.0517	2.76108	no
9.70496	such	in	3	1.52625	0.157265	no
20.5223	result	into	3	28.528	1.3901	yes, see table 7
69.9399	line	of	2	99.7009	1.42552	no
12.5842	issue	out	2	31.1721	2.47708	yes: A thesis will not be issued out of the Library
11.2238	influence	on	2	30.9885	2.76097	yes: There is also some political factor which also influences on cultural ...
74.3361	exciting	on	2	216.92	2.9181	no, almost identical sentence twice in same
9.33135	add	into	2	19.613	2.10184	no, parsing error

Table 8. Results on ICE India subpart, O / E ratio > 8 and O / E (BNC) < 3

We get fewer hits and fewer true positives, as the sparse data problem is considerably more acute. There is probably less structural nativisation in written texts than in spoken texts, which also contributes to the better results we get when including the spoken part. Repetitions of similar sentences also affect our findings. Interestingly, we also get two new findings.

3.6 Further verb-PP combinations

As mentioned, if we use less strict thresholds we get longer lists with much lower precision, but more instances are recalled. Going through longer lists lead to the following additional findings.

In ICE Fiji:

(21) Papua New Guinea where its Constitution emphasises on equal participation by women citizens (ICE-FJ:W1A-016)

(22) ... today the indigenous Fijians are still marginalised from the development process (ICE-FJ:W2B-012)

(23) The downloaded data was collated, analyzed and summarized into Table III. (ICE-FJ:W2A-033)

(24) I can't sleep from worrying. (ICE-FJ:W2F-017)

Example 24 contains an adjunct, which we have included in the automatic approach. Typically, collocations are complements, but many adjuncts can also be found, for example *sigh with relief, roar with laughter, appear before magistrate, prove beyond doubt*

(Lehmann and Schneider 2011).

In ICE India: Written only:

(25) Of course modern technology is improving the quality and hence even the hardened antagonists are switching over to them. (ICE-IND:W2D-019)

(26) You had the guts of your blighted mother to complain against us to the Governor. (ICE-IND:W2F-018)

(27) Wings are absent to apterygotes. (ICE-IND:W1A-019)

(28) The rule is that the company is the right person to sue and that it is not open to the individual members to assume to themselves the right of suing in the name of the company (ICE-IND:W2A-016)

Including Spoken:

(29) He was using the stones and preparing instruments out of it (ICE-IND:S1A-072)

(30) he has described all about that. (ICE-IND:S1A-092)

(31) the government of late has decided to slash down the export target for the year. (ICE-IND:S1B-056)

(32) ... retro-rockets were automatically fired to slow off the spaceship. (ICE-IND:S1B-006)

(33) he tried to enlighten the people and be aware towards all these irregularities. (ICE-IND:S1A-007)

We have also conducted experiments on verb-PP combinations that occur several times in ICE Fiji but are entirely absent in the BNC. The lists are dominated by adjuncts and by repetitions. The list of unseen verb-PP combinations that appear at least twice in ICE Fiji are given in table 9.

O / E ratio*	Head	Prep	f (Fiji)	manual comment
6273	strengthen	along	4	no, repetitions
3310	thump	up	2	no, repetitions
555	nest	around	2	no
330	download	by	2	no, internet age
1029	discriminate	since	2	no, repetitions
3463	cut	onto	2	no, repetitions
52	crosslink	with	2	no
2298	collaborate	with	2	no
117	choreograph	for	2	no
137	chirp	in	2	no
348	bag	from	2	no, parsing error

*A frequency of 0.1 was assumed for all unseen events, which makes it possible to calculate O / E for unseen combinations. Such smoothing techniques are standardly used in statistics.

Table 9. Verb-PP combinations unseen in the BNC while occurring at least twice in ICE Fiji

Including hapax legomena leads to a considerably longer lists, many false positives but also a few true positives, which are given in the following.

In ICE India: Written only:

(34) Adi Asenaca said an Asian Development Bank poverty participation survey listed down forms of poverty in the country and her ministry was following up on the recommendations. (ICE-FJ:W2C-013)

(35) Many still insist that they can get formal education due to insufficient funds and how to indulge into such activities where they get easy money and feed themselves. (ICE-FJ:W1A-020)

(36) Ravi watched horrified as his mother crashed towards the floor. (ICE-FJ:W2F-012)

(37) As a result some or nearly most women in the world have now turned to becoming prostitutes in order to cope up with poor living standard, they may be experiencing. (ICE-FJ:W1A-016)

4. Comparison of methods and conclusion

Based on the results that we obtained and our experiences with the processes of both the manual and the semi-automatic method, we compare advantages and disadvantages of each method in this section. The two methodological approaches to verb-preposition combinations both presented viable options with a number of results.

The particular advantages of the manual method described in section 2 of the present paper are the following: The analysis is very fine-grained, with high precision and recall. It is self-contained within each corpus under observation, which entails that phenomena can be detected for each variety without the need for a database for comparison. Furthermore, the method grants control over the scope of analysis; only manual analysis allows for a context-based semantic examination (see section 2.3). The disadvantages of the manual method, on the other hand, are first and foremost that it is very tedious and time-consuming. Therefore, it is technically barely possible to conduct it in connection with large corpora or highly frequent prepositions. Second,

this method relies on a predetermined starting point, i.e. a set of prepositions, as well as on codified and standardised dictionaries to assess the status of unrecorded verb-preposition combinations.

The semi-automatic method has advantages and disadvantages as well. A first advantage is that the method is corpus-driven, no prior set of prepositions needs to be assumed to start with, and theoretically, findings that are entirely different from those reported previously could be found. Second, the method scales well, not only to all prepositions, but also e.g. to adjective-PP combinations or to much larger texts. As sparse data is a serious issue, this method can only use its full potential when applied to much larger corpora. We will take a step into this direction in section 5. A first disadvantage of the semi-automatic method is that it misses many instances; it has relatively low recall, especially of semantically fine-grained distinctions. A second disadvantage is that manual interaction is still needed, the suggested results contain very many false positives. Furthermore, the method is particularly sensitive to duplicates, i.e. the same construction occurring several times in the same or a thematically related text.

We have partly found the same new verb-PP constructions using two diametrically opposed methods, and partly found different verb-PP constructions. Considering the different results, the methods complement each other precisely as they are very different in nature: they allow a researcher to attain much higher recall than either of the two methods on its own. The large overlap in results validates both approaches and gives one an assessment of the recall of each method.

5. Outlook: Scaling up with the *Statesman Corpus* (Semi-Automatic)

We are using existing ICE corpora for this pilot study, but the aim is to apply the same methodology to larger, web-derived (and thus somewhat 'messier') data. Using larger texts in the future will hopefully allow us to get a clearer distinction between production errors (slips of the tongue, typos, etc.) and structural nativisation phenomena: production errors remain nonce or rare occurrences, while structural nativisation phenomena report more hits. Concerning the semi-automatic method, we are using larger corpora, for example a subset of the Indian *The Statesman* newspaper and again compare to the BNC. A 3 million words excerpt of *The Statesman* Newspaper Archive excerpt has for example given us the findings listed in Table 10. We show the results obtained after manual filtering. We generally get more hits than on the small ICE corpora, but we also get many near-duplicates as newspaper articles may be repetitive. The results are also affected by genre differences: in comparison to the BNC, all or almost all texts come from the news genre.

verb	prep	f	Example
arrest	from	128	Seventeen contractual workers were arrested from the spot.
emphasise	on	12	He was a great reformer and throughout his life emphasised on the concepts of women's education and women's empowerment.
attach	with	8	... they would have to exercise caution in attaching themselves with projects they are not comfortable with.
aware	about	5	Tara Cancer Foundation, an NGO has been set up ... to make people aware about cancer ...
alert	about	5	However, we have to be alert about any possible attack .
aspire	for	4	It is Bollywood and not serious film makers that aspires too much for the Oscar glory these days.
list	out	4	I just can not understand your logic, he said and listed out statistics on funds allocated for various rural development projects.
discuss	about	3	The two also discussed about the entry of foreign educational institutions in India ...
dismiss	off	3	... but the South African had the last laugh by dismissing him off the last ball of the over.
devote	for	3	... Dr Ambedkar devoted his life for social justice for backward classes in the country.
rid	off	2	... and eight balls later Ajantha Mendis got rid off Sarwan ...
blind	into	2	the ... government in France had been blinded by supposed French interests in the region into siding with radical ... Hutu groups.

Table 10. New Verb-PP combinations found using the semi-automatic method on an excerpt of the *The Statesman* Newspaper Archive

Sources

Indian *The Statesman* newspaper: http://www.thestatesman.net/

Bibliography

Bautista, Maria Lourdes S. & Andrew B. Gonzales. 2006. "Southeast Asian Englishes." *The Handbook of World Englishes*, ed. by Braj B. Kachru, Yamuna Kachru & Cecil L. Nelson, 130–44. Malden, MA: Blackwell.

Biber, Douglas, Susan Conrad & Randi Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Bick, Eckhard. 2003. "A CG & PSG hybrid approach to automatic corpus annotation". *Proceedings of SProLaC2003*, ed. by Kiril Simow & Petya Osenova, 1–12. Lancaster: Lancaster University. http://www.bultreebank.org/SProLaC03Proceedings.html

Bick, Eckhard. 2010. "FrAG, a hybrid constraint grammar parser for French". *Proceedings of LREC 2010*, ed. by Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner & Daniel Tapias. Valletta: European Language Resources Association (ELRA).

Biewer, Carolin. 2009. “Modals and semi-modals of obligation and necessity in South Pacific Englishes”. *Anglistik* 20(2): 41–55.

Biewer, Carolin, Marianne Hundt & Lena Zipp. 2010. “How a Fiji corpus? Challenges in the compilation of an L2 ICE component.” *ICAME Journal* 34: 5–23

Carroll, John, Guido Minnen & Edward Briscoe. 2003. “Parser evaluation: using a grammatical relation annotation scheme”. *Treebanks: Building and Using Parsed Corpora*, ed. by Anne Abeillé, 299–316. Dordrecht: Kluwer.

Collins COBUILD Phrasal Verbs Dictionary. 2002. John Sinclair, ed. Glasgow: HarperCollins

Collins COBUILD Advanced Learner’s Dictionary (Resource Pack CD) - Lingea Lexicon. 2003. Glasgow: HarperCollins.

Evert, Stefan. 2009. “Corpora and collocations”. *Corpus Linguistics. An International Handbook*, article 58, ed. by Anke Lüdeling & Merja Kytö, 1212–1248. Berlin: Mouton de Gruyter.

Foley, Joseph A., ed. 1988. *New Englishes: The Case of Singapore.* Singapore: Singapore University Press.

Haverinen, Katri, Filip Ginter, Sampo Pyysalo & Tapio Salakoski. 2008. “Accurate conversion of dependency parses: targeting the Stanford scheme”. *Proceedings of Third International Symposium on Semantic Mining in Biomedicine (SMBM 2008)*, Turku, Finland, 2008.

Lehmann, Hans Martin, & Gerold Schneider. 2009. “Parser-Based Analysis of Syntax-Lexis Interaction”. *Corpora: Pragmatics and Discourse. Papers from the 29th International conference on English language research on computerized corpora (ICAME 29) (Language and computers 68)*, Ascona, Switzerland, 14–18 May 2008, ed. by Andreas H. Jucker, Daniel Schreier & Marianne Hundt, 477–502. Amsterdam: Rodopi.

Lehmann, Hans Martin – Gerold Schneider. 2011. “A large-scale investigation of verb-attached prepositional phrases”. *Methodological and Historical Dimensions of Corpus Linguistics*, ed. by Paul Rayson, Sebastian Hoffmann & Geoffrey Leech. (Studies in Variation, Contacts and Change in English 6). Helsinki: Research Unit for Variation, Contacts, and Change in English. http://www.helsinki.fi/varieng/journal/volumes/06/lehmann_schneider/

Mukherjee, Joybrato & Sebastian Hoffmann. 2006. “Describing verb-complementational profiles of New Englishes. A pilot study of Indian English.” *English World-Wide* 27(2): 147–173.

Mukherjee, Joybrato. 2009. “The lexicogrammar of present-day Indian English”. *Exploring the Lexis-Grammar Interface*, ed. by Ute Römer & Rainer Schulze, 117–135. Amsterdam: John Benjamins.

Nesselhauf, Nadja. 2009. “Co-selection phenomena across New Englishes. Parallels (and differences) to foreign learner varieties”. *English World-Wide* 30(1): 1–26.

van Noord, Gertjan & Gosse Bouma, 2009. “Parsed Corpora for Linguistics”. *Proceedings of the EACL 2009 Workshop on the Interaction between Linguistics and Computational Linguistics: Virtuous, Vicious or Vacuous?* Athens, Greece, 33–39.

Olavarría de Ersson, Eugenia, & Shaw, Philip. 2003. “Verb Complementation Patterns in Indian Standard English”. *English World-Wide* 24(2): 137–161.

Oxford English Dictionary Online. 2009. Oxford University Press. http://www.oed.com

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech & Jan Svartvik. 1985. *A Comprehensive Grammar of the English Language*. London: Longman.

Sand, Andrea. 2004. “Shared Morpho-syntactic features in contact varieties of English: Article use”. *World Englishes* 23(2): 281–298.

Schneider, Edgar W. 2004. “How to trace structural nativization: particle verbs in world Englishes”. *World Englishes* 23(2): 227–249.

Schneider, Edgar W. 2007. *Postcolonial English. Varieties around the world* (Cambridge Approaches to Language Contact). Cambridge: Cambridge University Press.

Schneider, Gerold. 2008. *Hybrid Long-Distance Functional Dependency Parsing*. Ph.D. dissertation, Institute of Computational Linguistics, University of Zurich.

Schneider, Gerold & Marianne Hundt. 2009. “Using a parser as a heuristic tool for the description of New Englishes.” *Proceedings of the Fifth Corpus Linguistics Conference*, Liverpool, 20–23 July 2009. http://ucrel.lancs.ac.uk/publications/cl2009/

Schreier, Daniel. 2003. *Isolation and Language Change: Contemporary and Sociohistorical Evidence from Tristan da Cunha English* (Palgrave Studies in Language Variation 1). Houndmills/Basingstoke & New York: Palgrave Macmillan.

Sedlatschek, Andreas. 2009. *Contemporary Indian English: Variation and Change*. Amsterdam & Philadelphia: John Benjamins.

Stubbs, Michael, 1995. “Collocations and semantic profiles: on the cause of the trouble with quantitative studies”. *Functions of Language* 2(1): 23–55.

Villavicencio, Aline. 2006. “Verb-Particle Constructions in the Wold Wide Web”. *Syntax and Semantics of Prepositions*, ed. by Patrick Saint-Dizier, 115–130. Dordrecht: Springer.

Xiao, Richard. 2009. “Multidimensional analysis and the study of world Englishes”. *World Englishes* 28(4): 421–450

Zipp, Lena. forthcoming. *Exo- and endonormative models in Fiji – A corpus-based study on the dynamics of first and second language varieties with a focus on Indo-Fijian English*. Ph.D. dissertation, English Department, University of Zurich.

Studies in Variation, Contacts and Change in English 13: Corpus Linguistics and Variation in English: Focus on Non-Native Englishes

Last updated 2013-05-15 by Joe McVeigh